Lau, C.L.; Mayfield, H.J.; Lowry, J.H.; Watson, C.H.; Kama, M.; Nilles, E.J.; Smith, C.S. (2017) [Accepted Manuscript] Unravelling infectious disease eco-epidemiology using Bayesian networks and scenario analysis: A case study of leptospirosis in Fiji. Environmental modelling & software. ISSN 1364-8152 DOI: https://doi.org/10.1016/j.envsoft.2017.08.004

DOI: 10.1016/j.envsoft.2017.08.004

Title: Unravelling Infectious Disease Eco-epidemiology using Bayesian Networks and Scenario Analysis: A Case Study of Leptospirosis in Fiji

Authors and affiliations:

Colleen Lau[a,b], Helen Mayfield[a], John Lowry[c,d], Conall Watson[e], Mike Kama[f], Eric Nilles[g], Carl Smith[h]

[a] Department of Global Health, Research School of Population Health, Australian National University, Canberra, ACT, Australia

[b] Child Health Research Centre, The University of Queensland, Brisbane, QLD, Australia

[c] School of People, Environment and Planning, Massey University, Palmerston North, New Zealand

[d] School of Geography, Earth Science and Environment, University of the South Pacific, Suva, Fiji

[e] Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

[f] Fiji Centre for Communicable Disease Control, Ministry of Health and Medical Services, Suva, Fiji

[g] Division of Pacific Technical Support, World Health Organization, Suva, Fiji

[h] School of Agriculture and Food Science, University of Queensland, Brisbane, Queensland, Australia

Corresponding author:

Dr Colleen Lau (MBBS, MPH&TM, PhD, FRACGP)

Department of Global Health, Research School of Population Health,

Australian National University

62 Mills Road, Canberra, ACT 0200, Australia

Email: colleen.lau@anu.edu.au. Phone: +61 402134878. Fax: +61 2 6125 5608.

ABSTRACT

Regression models are the standard approaches used in infectious disease epidemiology, but have limited ability to represent causality or complexity. We explore Bayesian networks (BNs) as an alternative approach for modelling infectious disease transmission, using leptospirosis as an example. Data were obtained from a leptospirosis study in Fiji in 2013. We compared the performance of naïve versus expert-structured BNs for modelling the relative importance of animal species in disease transmission in different ethnic groups and residential settings. For BNs of animal exposures at the individual/household level, R^2 for predicted versus observed infection rates were 0.59 for naïve and 0.75-0.93 for structured models of ethnic groups; and 0.54 for naïve and 0.93-1.00 for structured models of residential settings. BNs provide a promising approach for modelling infectious disease transmission under complex scenarios. The relative importance of animal species varied between subgroups, with important implications for more targeted public health control strategies.

KEYWORDS

Bayesian Networks, Infectious Diseases Epidemiology, Leptospirosis, Zoonoses, Environmental Health, Public Health

SOFTWARE AVAILABILITY

Name: Netica version 5.12

Developer: Norsys Software Corporation

Address: 3512 West 23rd Ave, Vancouver, BC, Canada

Tel: +1 604 221 2223. Email: info@norsys.com

Availability: www.norsys.com

DATA AVAILABILITY:

The data were collected from small communities in Fiji, and participants could potentially be re-identifiable if the study data were fully available, e.g. by diagnosis of leptospirosis, demographics, occupation, and household GPS locations. Public deposition of the data would compromise participant privacy, and therefore breach compliance with the protocol approved by the research ethics committees. For researchers who meet the criteria for access to confidential information, data can be requested via the Human Research Ethics Committee at the Australian National University. Email: human.ethics.officer@anu.edu.au. Phone: +61 (2) 6125 3427.

INTRODUCTION

The growing discipline of infectious disease eco-epidemiology seeks to understand the environmental, ecological, and socio-demographic drivers of emergence, transmission, and outbreaks.[1-3] The drivers depend on complex interactions between humans, the natural environment (e.g. climate and vegetation), the anthropogenic environment (e.g. urbanisation and land use), vectors (e.g. insects and animals), and carriers (e.g. water, soil, and air).[4] Regression models are the most common approaches to risk factor analysis in infectious disease epidemiology; while they are widely accepted and understood, there are important drawbacks when studying complex systems, and the need for more novel epidemiological approaches are being increasingly recognised.[5-10] Standard regression models rely on an explicit assumption of independence amongst the predictor variables as well as independence between units, which is often not true in the real world of disease transmission, and could potentially result in oversimplification of models. Standard regression models do not allow strongly correlated predictor variables to be retained, even if each variable might play crucial and distinct roles in transmission. Standard regression models therefore have limitations in their capacity to disentangle the intricate associations between risk factors, drivers, triggers, and outcomes.[7]

Causal models such as Bayesian networks (BNs) have the ability to represent causality as well as incorporate relationships between predictor/indicator variables, and may provide an alternative approach to more accurately model complex systems.[11,12] Other methods used to model complex systems and incorporate collinearity include the use of interactions in regression analysis, regression trees, structured equation models, path analysis and multilevel hierarchical models. Compared to these methods, Bayesian network models have added advantages of being both visually more intuitive and having interactive interfaces that can be used to assess complex scenarios and produce real-time outputs. In particular, the ability to define scenarios that include strongly correlated predictor variables is difficult to achieve with regression models. However, BNs also have certain limitations when modelling complex systems. BNs generally use discretised variables and produced outputs that are discrete outcomes or events, and discretisation of continuous variables is sometimes associated with loss of data resolution. Also, BNs are not dynamic and cannot incorporate feedback loops, a potentially important consideration for complex models.

Leptospirosis is an important zoonotic disease worldwide that causes an estimated one million severe cases per year, with particularly high risk in tropical and subtropical regions.[13,14] Humans are infected through direct contact with infected mammals (including rodents, livestock, pets, and wildlife), or contact with water or soil contaminated by urine of infected animals. Drivers of transmission are complex and include individual behaviour, socio-demographics, culture, lifestyle, contact with animals, and the natural environment.[15-17] Environmental drivers for leptospirosis transmission, emergence, and outbreaks are increasingly being recognised, raising concerns that transmission and flood-related outbreaks could intensify with global change in both natural and anthropogenic environments.[15,18,19] In developing countries, rapid population growth often results in urbanisation, slums, poor sanitation, poverty, subsistence livestock and agricultural intensification – all of which are important drivers of zoonotic disease transmission.[17,20] The Pacific Islands are particularly vulnerable to the health impacts of climate change because of all of the socio-demographic, geographic, and environmental factors mentioned above,[21,22] and leptospirosis causes significant health impact in the region.[23-28]

Over the past decades, Fiji has experienced increasing incidence and outbreaks of leptospirosis.[27,29-31] Two post-flooding outbreaks occurred in 2012, resulting in over 500 cases and 40 deaths. An eco-epidemiological study conducted in 2013 found a community leptospirosis seroprevalence (the percentage of a population with detectable leptospirosis antibodies in their blood) of 19·4% using the microscopic agglutination test (MAT), with significant variation between ethnic groups and residential settings. The findings of the study have been published, focusing on risk factor analysis using standard regression approaches.[27] The study provided important insights into leptospirosis eco-epidemiology in Fiji, but there remain multiple unanswered questions with important public health implications. Important questions regarding the reasons for the disparate risk between ethnic groups and residential settings have not been clearly answered, but it is possible that niche-specific interventions may be required for more effective public health control measures. For example, intervention strategies may need a different focus for each ethnic group and/or vary between urban, peri-urban, and rural areas. The study also raised questions about the relative importance of animal species in human infections, a fundamental question when prioritising public health interventions for leptospirosis. On univariate regression analysis, infection was associated with contact with multiple animal species, including rodents, mongoose, dogs, and multiple species of livestock. However, there were significant correlations between presence of different animals species (e.g. people who own pigs are also more likely to own cows), and on multivariable regression analyses, the only animal-related predictor variables retained in the final model were the presence of pigs in the community and high cattle density. Based on these results, can we assume that animal species other than pigs and cattle did not play an important role in human infections? Or could other species be also important, but excluded from multivariable regression models because they were strongly correlated with exposure to pigs or cattle? Also, might the relative importance of different animal species differ between ethnic groups and residential settings, and therefore require more tailored interventions? These questions highlight some of the limitations of using standard regression analysis to model infectious diseases with complex transmission dynamics and environmental drivers.

In this paper, we explore the use of BNs as an alternative methodological approach for modelling the eco-epidemiology of infectious diseases, using leptospirosis in Fiji as a case study. Firstly, the study aims to improve model performance of BNs by building models that better represent and explain causality. Secondly, the study aims to use BNs to determine the relative importance of animal species in disease transmission in different ethnic groups and residential settings.

MATERIALS and METHODS

Study location and setting

Fiji has a population of 837,217 [32] living in urban, peri-urban, and rural settings in tropical islands. Two main ethnic groups, iTaukei (indigenous Fijian) and Indo-Fijians (Fijians of Indian descent), account for 57% and 35% of the population respectively.[32] Subsistence livestock are common in backyards and communal areas, particularly in rural areas. Rodents, mongoose, dogs, and cats are abundant in both urban and rural areas.

Data sources

This study used a database from a recently published study of leptospirosis in Fiji, which was designed to include a representative sample of the country's population.[27] Briefly, the cross-sectional community seroprevalence study included 2,152 participants aged 1 to 90 years from 81 communities on the three main islands of Fiji. Blood samples were collected from each participant, and the microscopic agglutination test (MAT) was used to determine the presence of *Leptospira* antibodies, an indicator of previous infection. Data on socio-demographics, environmental factors, and animal exposure were obtained from questionnaires, population census, agricultural census, World Bank poverty survey, and geo-referenced environmental data. Data were linked to household locations using geographic information systems (GIS) to generate a richly structured geospatial database that relates risk factors and outcome (presence of *Leptospira* antibodies) for each individual.

Predictor/indicator variables examined in this study

In this study, we focused on more in-depth analysis of the following predictor/indicator variables, and built scenarios related to animal exposure in different ethnic groups and residential settings:

- Ethnic group:
  - iTaukei, Indo-Fijian (other ethnic groups were excluded because they accounted for only 2% of the study population)
- Residential setting:
  - Urban, peri-urban, rural
- Exposure to animals at the individual/household levels:
  - Physical contact with rodents and/or mongoose
  - Dogs, cats, chickens, pigs, cows, goats, horses
- Exposure to animals at the community level:
  - Pigs, cows, goats, horses

Table 1 provides a summary of the distribution of ethnic groups and residential settings in the study population, and the variations in *Leptospira* seroprevalence found in the 2013 study.

Table 1. Summary of distribution of ethnic groups and residential settings in dataset, and differences in observed seroprevalence in each subgroup.

Variable	Number of subjects	% of total subjects	Observed seroprevalence	Univariate odds ratio (regression analysis)	*p* value
Total sampled	2152	100%	19.4%		
Ethnic groups					
Indo-Fijian	459	21.3%	7.4%	1	
iTaukei	1651	76.7%	22.7%	3.66	<0.001
Other	39	2.0%	20.5%	3.23	0.114
Residential settings					
Urban	579	26.9%	11.1%	1	
Peri-urban	287	13.3%	15.3%	1.46	0.074
Rural	1286	59.8%	24.0%	2.54	<0.001

Adapted from Lau et al 2016 (27)

The frequency of exposure to animal species in each ethnic group and residential setting were summarised. For individual/household-level analyses, physical contact with rodents or mongoose were included in the analyses but mere sighting of these species around the home were not included because 85.9% and 77.1% of participants reported sighting of rodents and mongoose respectively; these variables therefore did not provide good discriminatory power and were not statistically associated with the presence of *Leptospira* antibodies at a univariate level. Similarly, the presence of rodents, mongoose, dogs, cats and chickens were not assessed at the community level because these species were ubiquitous.

Bayesian Networks

BNs are probabilistic models based on Bayes' theorem of conditional probability, composed of: i) directed acyclic graphs (DAGs) with nodes that represent variables and outcomes and arrows that define dependency between nodes, and ii) node probability tables (NPT).[33] BNs were constructed using Netica software.[34] Figure 1 shows a simple BN, where 'presence of *Leptospira* antibodies' (child node) is dependent on 'pigs in community' and 'residential setting' (parent nodes). 'Pigs in community' is in turn dependent on 'residential setting'. For child nodes that conditionally depend on their parent nodes, the NPT is called a conditional probability table (CPT) that defines the probabilistic relationship between the nodes. The CPT for 'Presence of *Leptospira* antibodies' (Table 2) shows that for a rural setting with pigs, there is a 27.5% probability of the presence of antibodies. For parentless nodes, e.g. 'residential setting', an unconditional probability table stores the prior probabilities of each state: e.g. Figure 1a shows that 59.8% of the population live in rural areas.

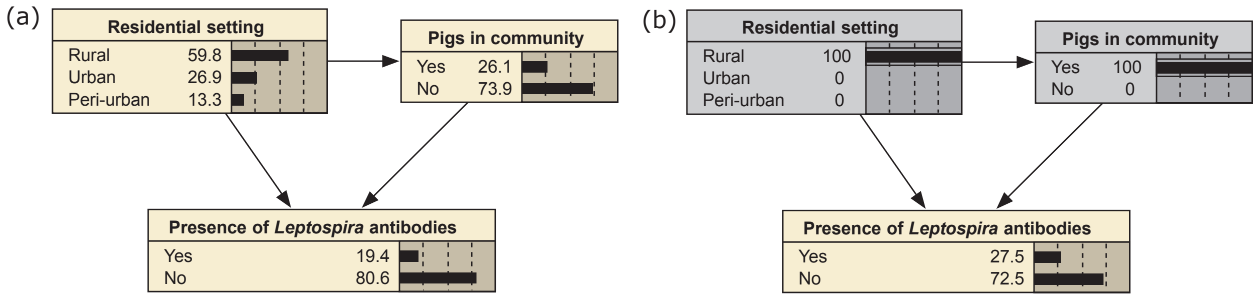

Figure 1. A simple Bayesian network for estimating the probability of the 'Presence of *Leptospira* antibodies' based on the presence/absence of pigs in the community and type of residential setting. The network has two predictor or "parent" nodes ('Pigs in community' and 'Residential setting') linked to the outcome or child node ('Presence of Leptospira antibodies'). The presence/absence of 'Pigs in community' is also dependent on 'Residential setting'. The 'Pigs in community' node includes two categories or 'states': Yes or No. The 'Residential setting' variable includes three states: Rural, Urban, and Peri-urban. In Figure 1a), the nodes were set to show the 'default probabilities' in the belief bars, which provide a reflection of the data, i.e. approximately 26.1% of the study population had pigs in their community, 59.8% lived in rural areas, and Leptospira antibodies were present in 19.4%. In Figure 1b), a scenario was defined by selecting belief bars to show that in a rural residential setting where pigs were present, the probability of *Leptospira* antibodies being present was 27.5%.

Table 2. Conditional probabilities table (CPT) for the 'Presence of *Leptospira* antibodies' node, showing the probabilities of the presence/absence of *Leptospira* antibodies for all combinations of states in the parent nodes ('Residential setting' and 'Pigs in community')

States of parent nodes		Probability of the Presence of *Leptospira* antibodies (%)	
Residential setting	**Pigs in community**	**Yes**	**No**
Rural	Yes	27.5	72.5
Rural	No	22.3	77.7
Urban	Yes	23.8	76.2
Urban	No	8.9	91.1
Peri-urban	Yes	25.9	24.1
Peri-urban	No	12.9	87.1

In naïve BNs, predictor/indicator variables are assumed to be independent. In structured BNs, causal dependencies between nodes can be defined using arrows, and each node can be used as predictor or indicator depending on the direction of the arrow. The graphical interface of BNs allows users to define scenarios by selecting states for each node (e.g. a rural community with pigs). When a node state is selected (referred to as inserting findings or evidence), the probabilities in all other nodes are updated using Bayes' Theorem of conditional probabilities according to the causal dependencies among nodes (probability propagation). NPTs and causal dependency can be learnt directly from data via parameter and structural learning algorithms, or derived from expert opinion.

Model structure and parameterisation

Three groups of BNs, one naïve and two expert-structured, were built and used to analyse scenarios of animal exposure for the two ethnic groups (iTaukei and Indo-Fijian) and three residential settings (urban, peri-urban, rural). Group A BNs were naïve networks, which assumed that all predictor/indicator variables were independent. Group B and C BNs were structured networks designed specifically to examine the role of each animal species in disease transmission in different ethnic groups and residential settings. BNs in Groups A, B, and C were compiled based on the influence diagrams in Figure 2. Table 3 shows the codes of the three groups of BNs for ease of reference.

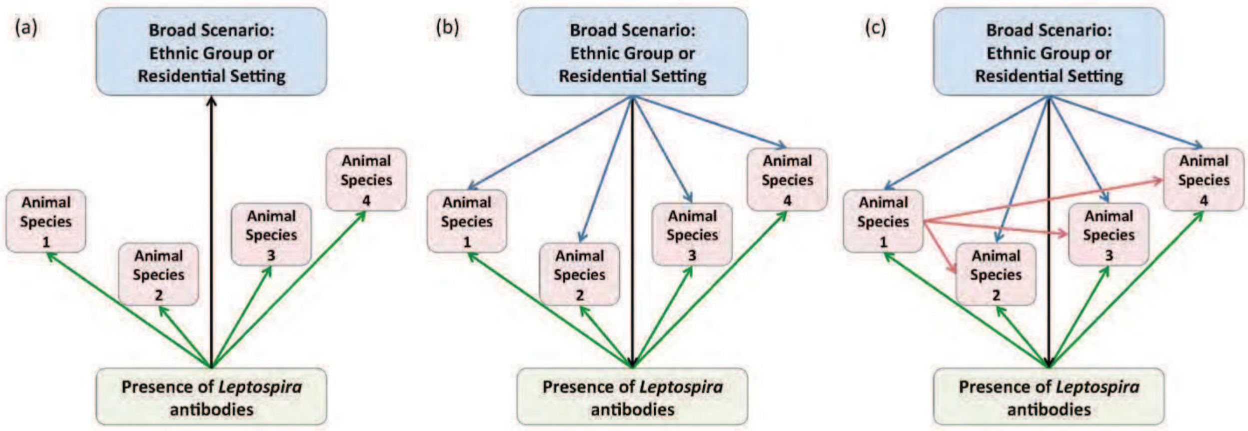

Figure 2. Frameworks for influence diagrams for a) Group A BNs were naïve networks and assume that all indicator variables were independent, with each variable individually linked to the outcome; b) Group B BNs were structured networks, and reflect that the broad scenario is a *predictor* (parent node) of the presence of each animal species (blue arrows), and each animal species is in turn an *indicator* (child node) of the outcome (green arrows); c) Group C were structured to also take into account interdependence between nodes related to animal exposure by creating links from species A to species B, C and D (red arrows). The broad scenario was also directly linked to the outcome (black arrow) to take into account the alternate exposure pathways (other than animal exposure) through which ethnicity and residential setting could influence infection risk (e.g. behaviour, occupation).

The influence diagram for Group A BNs (Figure 2a) assumes that all indicator variables were independent, and each variable was individually linked to the outcome (presence of *Leptospira* antibodies). The influence diagram for Group B BNs (Figure 2b) was structured to reflect that the broad scenario (ethnic group or residential status) is a *predictor* (parent node) of the presence of each animal species in the community (blue arrows), and each animal species is in turn an *indicator* (child node) of the presence of *Leptospira* antibodies (green arrows). Animal species nodes were not used as predictors of the outcome because this structure would have resulted in a very large conditional probability table for the outcome node, and undefined probabilities for a significant number of scenarios. It is more logical to have arrows pointing from cause to effect, but in some cases, the directions of arrows are reversed to avoid large conditional probability tables that are difficult to parameterise with available data. Reversing the direction of arrows is possible in a BN because inference can work both directions.[35] However, biological plausibility needs to be considered when determining the direction of causation, which is not necessarily the same as the direction of the arrows. For example, in our models, exposure to animals 'causes' an increased risk of leptospirosis, and not vice versa.

BNs in Group C (Figure 2c) were structured to also take into account dependence between the variables related to animal exposure. Links were created between the most common animal species and all other species (red arrows), resulting in conditional probabilities that take into account dependence between the animal variables, e.g. the presence of cows is correlated with the presence of goats, pigs, and horses. For BNs related to individual/household-level animal exposure, animal species were categorised into three groups: feral (rodents and mongoose), pets (dogs and cats) and livestock (goats, pigs, horses and cows). Dependencies were modelled only within each of the three animal groups. The broad scenario node was also directly linked to the outcome node (black arrow) to take into account the alternate exposure pathways (other than animal exposure) through which ethnicity or residential setting could influence infection risk (for example behaviour, occupation, poverty or sanitation).

Conceptually, Group A BNs are similar to standard regression models, where all predictor/indicator variables are independent. Group B BNs were structured to provide a better representation of the causal relationships between variables. Group C also considered interdependence between the animal variables. Unlike standard regression models, BNs are capable of incorporating and retaining strongly correlated variables in the final models, such as exposure to multiple animal species.

Model training and testing

Bayesian networks are driven by the Bayes theorem of conditional probability and allows prior knowledge to be incorporated into model predictions. Bayes theorem (Equation 1) states that the conditional probability of a hypothesis (H) occurring given evidence (E), can be calculated as the product of the probability of H and the conditional probability of E given H, divided by the probability of E.

$$P(H \mid E) = P(H) \times P(E \mid H) / P(E) \qquad \text{Equation 1}$$

In a BN, this formula is used to calculate and update conditional probabilities of all node states when evidence is inserted into one or more nodes. Probabilities for NPTs (including CPTs) can be either learnt from the data during model training, or defined by experts.

Networks were trained using the Expectation Maximisation algorithm[36] in Netica, and tested using two methods:

1. Model discrimination ability was measured using the area under the curve of the receiver operating characteristic (AUC). The AUC for each BN was calculated using trials, where 50% of the data were used to train the BN and populate the CPTs, and the other 50% used to test the BN (to

determine the accuracy of the predicted prevalence values). For each BN, repeated random subsampling was used to conduct 30 trials, and the average AUC reported.

2. Model calibration (measure of how well the model fits the data, or model goodness-of-fit) was measured by comparing predicted and observed probabilities for each set of BNs. For this purpose, BNs were trained using 100% of the dataset. The agreement between predicted probabilities of the presence of *Leptospira* antibodies under different scenarios and the observed probabilities (empirical data from the 2013 field study) were measured using R^2 and mean squared error (MSE). We examined scenarios based on ethnicity, residential location, and exposure to animal species. After defining a broad scenario of ethnicity or residential location, more specific scenarios of animal exposure were examined. We analysed the influence of each animal species individually, and also combinations of two and three animal species if these scenarios were reported by >3% of at least one ethnic or residential subgroup. Less common scenarios were not assessed because of insufficient data for robust predictions, and their low relevance for understanding disease transmission and informing public health interventions. Nodes that were not included in a scenario were left in their default state. Each trio of Group A, B, and C BNs were compared to determine whether predictive performance of models improved by structures that better represented causality.

Relative importance of animal species under different scenarios

The relative importance of each animal species in leptospirosis transmission for each ethnic group and residential setting were examined using the Group C BNs. To ascertain whether exposure to one or more animal species had a significant effect on seroprevalence, a test of proportions was used to determine if differences in predicted seroprevalence between exposed and unexposed groups were statistically significant at $p < 0.05$.

RESULTS

Bayesian network models

Based on the influence diagrams in Figure 2, 12 BNs were compiled. Differences between the BNs are summarised in Table 3, and each of the BNs were assigned a code for ease of reference. The structures and variables included in each set of BNs are shown in Figure 3A to 3D. The 'belief bars' in the figures show the probability distributions for the states of each node captured by the dataset, and reflect conditional probabilities between all connected nodes, e.g. Figure 3B shows that 76.8% of the study population are of iTaukei ethnicity, 26.1% reported the presence of pigs in the community, and 19.4% were seropositive for leptospirosis.

Table 3. Summary of the three groups of BNs used to examine the role of animal species in different ethnic groups and residential settings, and the codes used for each BN for ease of reference.

	Group A	Group B	Group C
Influence diagram	Figure 2a	Figure 2b	Figure 2c
Model type	Naïve Bayesian network	Structured Bayesian network	Structured Bayesian network
Assumptions about predictor/indicator variables	All predictor/indicator variables independent	Variables related to animal exposure were independent, e.g. presence of cows was not correlated with presence of other animal species.	Considered dependence between variables related to animal exposure, e.g. presence of cows was associated with the presence of other animal species.
Model structure	Each predictor/indicator variable individually linked to the outcome. Conceptually similar to regression models.	The broad scenario (ethnic group or residential status) was used as a *predictor* (parent node) of the presence of each animal species (blue arrows), and each animal species was in turn used as an *indicator* (child node) of the presence of *Leptospira* antibodies (green arrows). The broad scenario also directly linked to the outcome node (black arrow) to take into account the alternate exposure pathways (other than animal exposure) through which ethnicity or residential setting could influence infection risk (for example behaviour, occupation, poverty or sanitation).	In addition to the model structures for Group B BNs, Group C BNs also considered dependence between the variables related to animal exposure. Links were created between the most common animal species and all other species (red arrows), resulting in conditional probabilities that take into account dependence between animal variables, e.g. the presence of cows is correlated with the presence of other animal species. For BNs related to individual/household-level animal exposure, animal species were categorised into three groups: feral (rodents and mongoose), pets (dogs and cats) and livestock (goats, pigs, horses and cows). Dependencies were modelled only within each of the three animal groups.
Codes for BNs used to examine Ethnicity and Individual/household-level animal exposure (Figure 3A)	EI-A	EI-B	EI-C
Codes for BNs used to examine Ethnicity and Community-level animal exposure (Figure 3B)	EC-A	EC-B	EC-C
Codes for BNs used to examine Residential setting and Individual/household-level animal exposure (Figure 3C)	RI-A	RI-B	RI-C
Codes for BNs used to examine Residential setting and Community-level animal exposure (Figure 3D)	RC-A	RC-B	RC-C

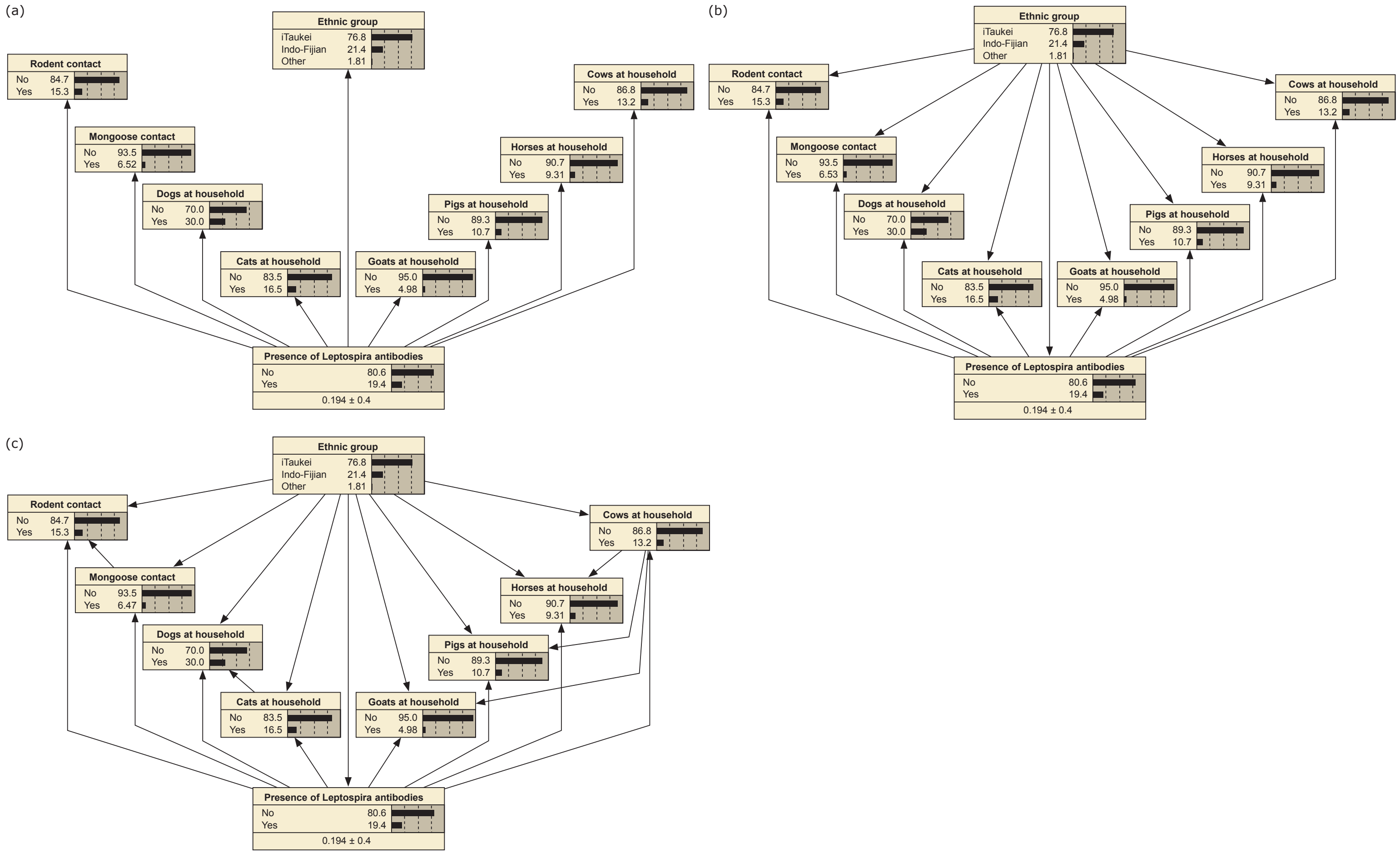

Figure 3A. BNs used to model the probability of the presence of *Leptospira* antibodies based on ethnicity and individual/household-level exposure to livestock animal species. a) BN EI-A, a naïve network assuming that all variables were independent, b) BN EI-B, a structured network that provides a better representation of interrelationships between variables, but assuming that animal variables were independent, and c) BN EI-C, structured network taking into account interdependence between animal variables.

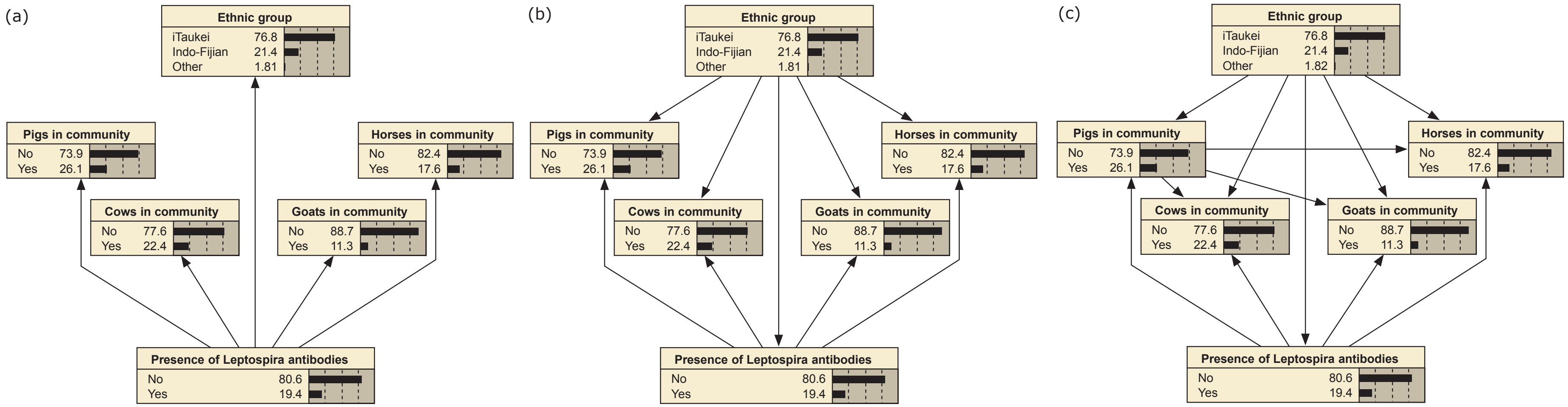

Figure 3B. BNs used to model the probability of the presence of *Leptospira* antibodies based on ethnicity and the presence of livestock animal species in the community: a) BN EC-A, a naïve network assuming that all variables were independent, b) BN EC-B, a structured network that provides a better representation of interrelationships between variables, but assuming that animal variables were independent, and c) BN EC-C, structured network taking into account interdependence between animal variables.

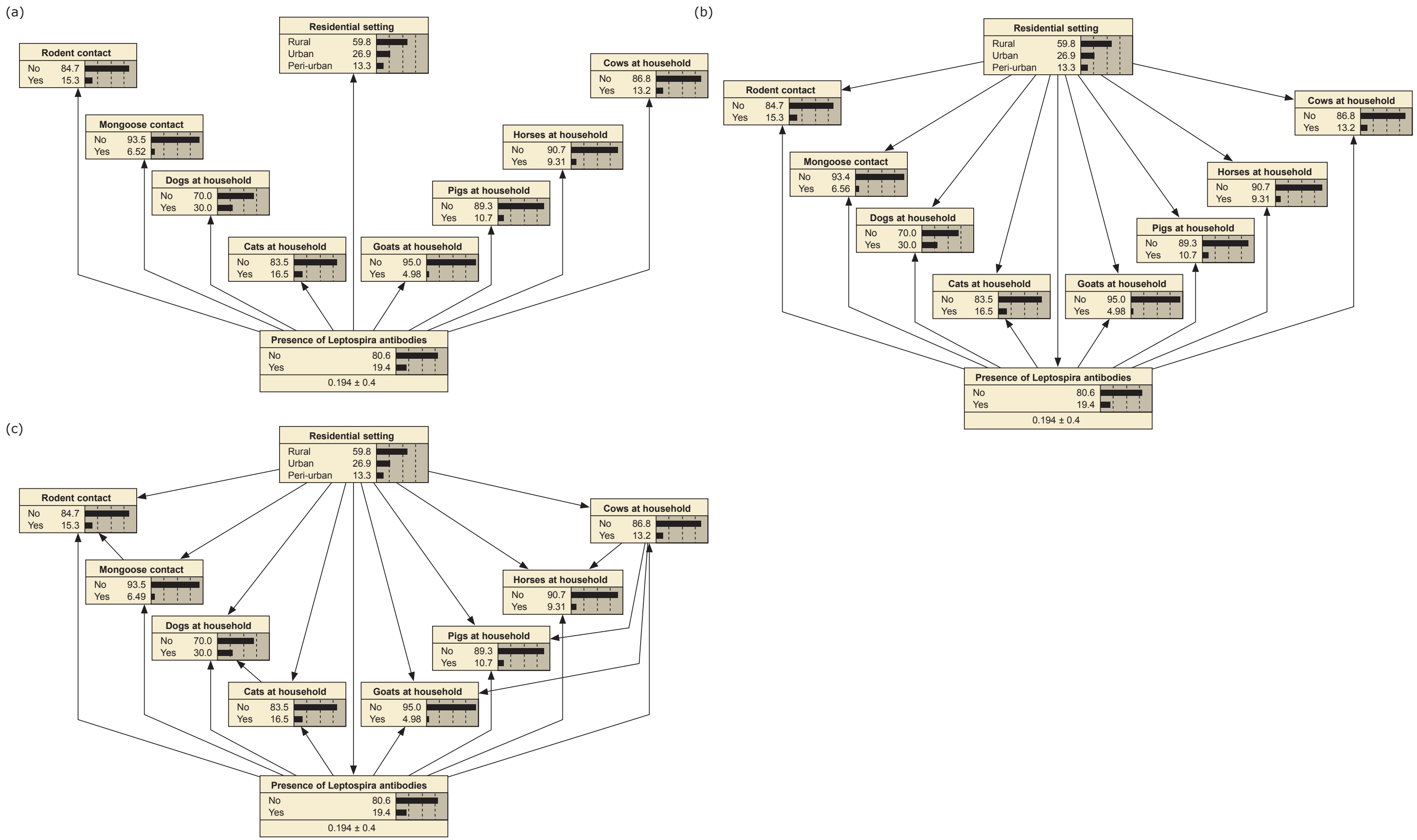

Figure 3C. BNs used to model the probability of the presence of *Leptospira* antibodies based on residential setting and individual/household level exposure to livestock animal species. a) BN RI-A, a naïve network assuming that all variables were independent, b) BN RI-B, a structured network that provides a better representation of interrelationships between variables, but assuming that animal variables were independent, and c) BN RI-C, structured network taking into account interdependence between animal variables.

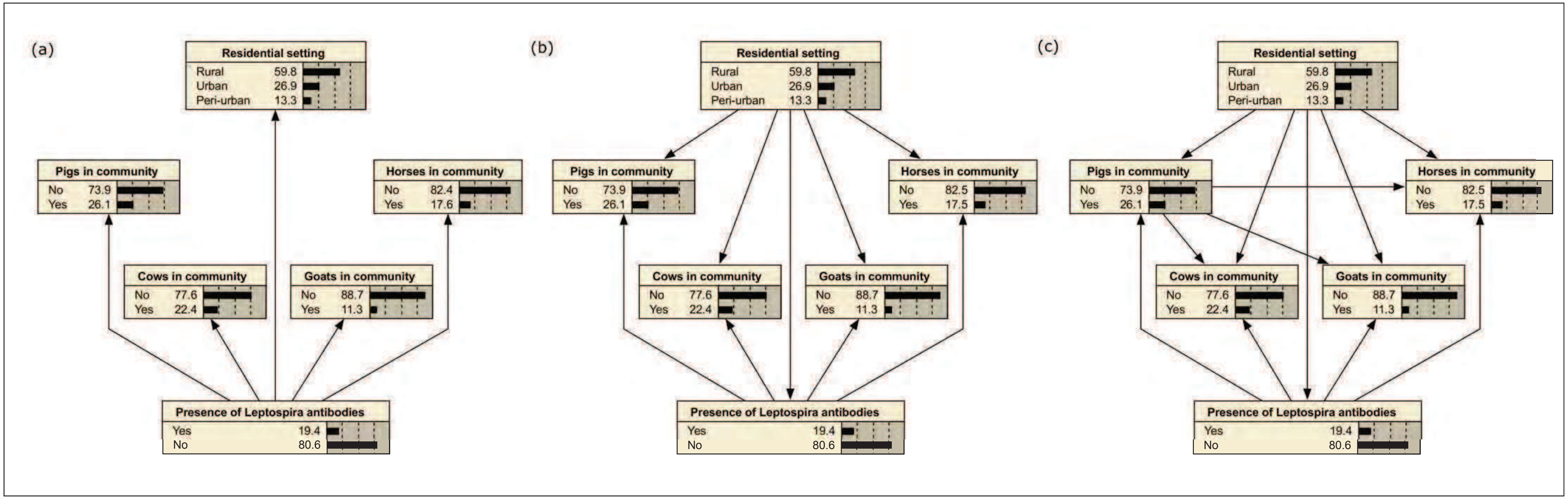

Figure 3D. BNs used to model the probability of the presence of *Leptospira* antibodies based on residential setting and the presence of livestock animal species in the community. a) BN RC-A, a naïve network assuming that all variables were independent, b) BN RC-B, a structured network that provides a better representation of interrelationships between variables, but assuming that animal variables were independent, and c) BN RC-C, structured network taking into account interdependence between animal variables.

Model testing

a) Model discrimination ability – AUC

The median AUC results over the 30 trials for each of the 12 BNs ranged from 0.59-0.61 (Table 4), and indicate poor (but better than random) model discriminatory ability. There were no significant differences in AUCs between Groups A, B and C BNs.

Table 4. AUC results over 30 trials for Group A, B, and C BNs.

Bayesian Network Code	Median AUC	Interquartile Range
Ethnicity and Individual/household-level exposure to animals:		
EI-A	0.61	0.60-0.61
EI-B	0.60	0.59-0.61
EI-C	0.59	0.58-0.60
Ethnicity and Community-level exposure to animals:		
EC-A	0.61	0.60-0.62
EC-B	0.61	0.60-0.62
EC-C	0.60	0.61-0.63
Residential setting and Individual/household-level exposure to animals:		
RI-A	0.61	0.60-0.61
RI-B	0.60	0.58-0.60
RI-C	0.59	0.58-0.60
Residential setting and Community-level exposure to animals:		
RC-A	0.61	0.61-0.62
RC-B	0.60	0.60-0.61
RC-C	0.60	0.59-0.61

b) Model calibration – predicted versus observed seroprevalence

Tables 5 to 8 show the scenarios of animal exposure for ethnic group and residential setting where at least 3% of one or more subgroups reported the exposure scenarios; these scenarios were included in further analyses. The tables also show the percentage of each subgroup that reported the animal exposures. For example, Table 6 shows the most common scenarios of community-level animal exposure(s) for each ethnic group, where at least 3% of one or more ethnic group reported that combination of animal exposure. Sections A, B, and C list the scenarios related to exposure to each animal species, combinations of two animal species, and combinations of three animal species respectively. If a scenario was reported by <3% of a subgroup, the predicted seroprevalence is not reported.

For each scenario of animal exposure shown in Tables 5 to 8, the predicted seroprevalence were calculated using the associated BNs and compared to the observed seroprevalence. For example, BNs EC-A, EC-B, and EC-C were used to predict seroprevalence for each of the scenarios of ethnicity and community-level animal exposure(s) shown in Table 6. Section B of Table 6 shows that 16.7% of iTaukei and 4·4% of Indo-Fijians reported the presence of both cows *and* horses in their community. And in iTaukei who reported exposure to both cows and horses, the observed seroprevalence was 25.5%, while the predicted seroprevalence using EC-A, EC-B, and EC-C were 36.3%, 29.4%, and 27.3% respectively.

Agreement between predicted and observed seroprevalence were measured using R^2 and MSE, and the correlations for each trio of Group A, B, and C models are shown in Figures 4 and 5. The figures show that R^2 values improved from 0.59 for EI-A to 0.93 for EI-C; 0.78 for EC-A to 0.93 for EC-C; 0.54 for RI-A to 1.00 for RI-C; and 0 for RC-A to 0.75 for RC-C. Similarly, MSE showed that Group C models produced the best agreement between predicted and observed seroprevalence. MSE were 67.1, 22.6 and 3.6 for EI-A, EI-B, and EI-C; 95.0, 67.2, and 7.1 for EC-A, EC-B, and EC-C; 46.8, 6.3, and 0.3 for RI-A, RI-B, and RI-C; and 144.8, 364.3, and 16.6 for RC-A, RC-B, and RC-C respectively. For each trio of BNs, the best predictive accuracy (highest R^2 and lowest MSE) was seen with Group C models.

Table 5. The most common individual/household-level exposure to animal species in each ethnic group. For rodents and mongoose, exposure was defined as physical contact with these animals. For other animal species, exposure was defined as presence of the animal species at the individual's household. BNs EI-A, EI-B and EI-C were used to predict seroprevalence under each of the scenarios shown below, and summarised in Figure 4a.

Section	Physical contact		Animal species present at household						% of population exposed to animal species		Observed seroprevalence* (%)		Predicted seroprevalence using EI-A (%)		Predicted seroprevalence using EI-B (%)		Predicted seroprevalence using EI-C (%)	
	Rodents	Mongoose	Dog	Cat	Cow	Goat	Horse	Pig	iTaukei n=1651	Indo-Fijian n=459	iTaukei	Indo-Fijian	iTaukei	Indo-Fijian	iTaukei	Indo-Fijian	iTaukei	Indo-Fijian
A. Exposure scenarios related to EACH animal species	X								17.3	6.3	27.3	6.9	30.0	10.5	27.8	7.1	27.9	7.1
		X							7.5	2.0	30.1	0.0	33.7	-	30.5	-	30.6	-
			X						26.0	43.6	24.0	9.5	22.7	7.42	24.1	9.5	24.1	9.5
				X					14.4	23.3	19.4	9.4	19.3	6.11	19.4	9.3	19.4	9.3
					X				14.1	10.7	27.0	18.4	29.7	10.3	27.1	18.4	27.1	18.4
						X			3.1	12.2	23.5	17.9	24.0	7.95	23.5	17.8	23.5	17.8
							X		10.5	5.7	27.0	19.2	30.0	10.5	27.0	19.2	27.0	19.2
								X	13.7	0.4	25.7	-	30.1	-	25.7	-	25.7	-
B. Exposure scenarios related to combinations of TWO animal species	X	X							4.6	1.1	29.0	0.0	42.6	-	36.6	-	29.7	-
			X	X					6.8	17.2	19.5	11.4	19.3	6.12	20.7	11.9	19.5	11.4
			X			X			1.5	8.1	24.0	21.6	-	7.95	-	22.2	-	22.2
			X		X				6.2	7.8	31.1	19.4	29.7	10.4	28.7	22.8	28.7	22.8
					X		X		7.2	2.6	28.6	16.7	38.2	-	31.9	-	28.6	-
C. Exposure scenarios related to combinations of THREE animal species			X	X		X			0.7	4.8	8.3	18.2	-	6.56	-	26.9	-	25.8
			X	X	X				1.8	3.9	13.8	27.8	-	8.58	-	27.5	-	26.5
			X		X	X			1.3	4.4	18.2	25.0	-	11.1	-	44.5	-	28.2
			X		X		X		3.5	2.6	29.3	16.7	38.2	-	33.7	-	30.3	-

*Overall observed seroprevalence in 2013 field study was 22.7% in iTaukei and 7.4% in Indo-Fijians. Predicted seroprevalence were only calculated for animal exposure scenarios reported by >3% of at least one subgroup; "-" indicates scenarios where predicted seroprevalence were not calculated.

Table 6. The most common community-level exposure to animal species in each ethnic group. Exposure was defined as the presence of the animal species at the individual's community. BNs EC-A, EC-B and EC-C were used to predict seroprevalence under each of the scenarios shown below, and summarised in Figure 4b.

Section	Animal species present in community				% of population exposed to animal species		Observed seroprevalence* (%)		Predicted seroprevalence using EC-A (%)		Predicted seroprevalence using EC-B (%)		Predicted seroprevalence using EC-C (%)	
	Cow	Goat	Horse	Pig	iTaukei n=1651	Indo-Fijian n=459	iTaukei	Indo-Fijian	iTaukei	Indo-Fijian	iTaukei	Indo-Fijian	iTaukei	Indo-Fijian
A. Exposure scenarios related to EACH animal species	X				24.8	12.6	25.6	12.1	28.6	9.88	25.6	12.1	25.6	12.1
		X			11.3	11.3	28.9	11.5	29.2	10.1	28.9	11.5	28.9	11.5
			X		21.1	5.4	26.2	16.0	29.4	10.2	26.2	16.0	26.2	16.0
				X	32.7	1.3	25.9	16.7	30.8	-	26.1	-	26.1	-
B. Exposure scenarios related to combinations of TWO animal species	X	X			10.7	6.8	30.7	9.7	36.0	13.3	32.3	18.3	29.7	14.0
	X		X		16.7	4.4	25.5	10.0	36.3	13.5	29.4	24.6	27.3	18.5
	X			X	18.6	0.9	26.4	0.0	37.9	-	29.3	-	26.4	-
		X	X		9.6	4.1	29.1	15.8	36.9	13.8	32.9	23.7	30.3	17.8
		X		X	9.5	0.7	29.3	33.3	38.5	-	32.8	-	29.3	-
			X	X	15.9	0.4	27.0	0.0	38.8	-	29.9	-	27.0	-
C. Exposure scenarios related to combinations of THREE animal species	X	X	X		9.4	3.5	29.5	6.3	44.5	18.0	36.6	34.7	30.7	13.0
	X	X		X	9.1	0.4	30.5	0.0	46.1	-	36.5	-	29.7	-
	X		X	X	13.4	0.4	26.2	0.0	46.5	-	33.4	-	27.3	-
		X	X	X	8.4	0.4	29.7	0.0	47.1	-	37.1	-	30.3	-

*Overall observed seroprevalence in 2013 field study was 22.7% in iTaukei and 7.4% in Indo-Fijians. Predicted seroprevalence were only calculated for animal exposure scenarios reported by >3% of at least one subgroup; "-" indicates scenarios where predicted seroprevalence were not calculated.

Table 7. The most common individual/household-level exposure to animal species in each residential setting. For rodents and mongoose, exposure was defined as physical contact with these animals. For other animal species, exposure was defined as presence of the animal species at the individual's household. BNs RI-A, RI-B and RI-C were used to predict seroprevalence under each of the scenarios shown below, and summarised in Figure 5a.

Section	Physical contact		Animal species present at household						% of population exposed to animal species			Observed seroprevalence* (%)			Predicted seroprevalence using RI-A (%)			Predicted seroprevalence using RI-B (%)			Predicted seroprevalence using RI-C (%)		
	Rodents	Mongoose	Dog	Cat	Cow	Goat	Horse	Pig	Urban n=579	Peri-urban n=287	Rural n=1286	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural
A. Exposure scenarios related to EACH animal species	X								13.6	12.9	16.1	16.5	27.0	29.0	15.4	20.9	31.6	16.3	26.9	29.5	16.3	26.9	29.5
		X							4.0	4.2	7.8	17.4	25.0	32.0	17.7	23.9	35.4	17.1	25.4	32.6	17.2	24.9	32.6
			X						26.6	33.4	30.7	12.3	13.5	23.5	11.1	15.3	24.0	12.3	13.5	23.6	12.3	13.5	23.6
				X					15.2	22.3	15.8	5.7	14.1	21.7	9.2	12.8	20.5	5.7	14.0	21.7	5.7	14.0	21.7
					X				3.5	5.2	19.4	25.0	6.7	26.9	15.2	20.7	31.3	25.0	6.6	27.0	25.0	6.6	27.0
						X			0.9	1.0	7.7	22.0	0.0	21.2	-	-	25.4	-	-	21.2	-	-	21.2
							X		1.4	2.1	14.5	0.0	16.7	27.4	-	-	31.6	-	-	27.5	-	-	27.5
								X	3.6	6.6	14.8	33.3	26.3	25.3	15.4	21.0	31.7	33.3	26.2	25.3	33.3	26.2	25.3
B. Exposure scenarios related to combinations of TWO animal species	X	X							2.8	3.1	4.6	18.8	33.3	30.5	-	31.5	44.5	-	40.8	39.0	-	33.2	31.4
			X	X					5.4	14.3	9.8	3.2	17.1	19.1	9.2	12.8	20.5	6.4	12.3	21.3	3.2	17.0	19.1
			X		X				0.7	2.4	10.0	50.0	0.0	28.9	-	-	31.3	-	-	26.5	-	-	26.5
					X		X		0.7	1.7	9.5	0.0	0.0	29.5	-	-	40.0	-	-	30.6	-	-	29.6
C. Exposure scenarios related to combinations of THREE animal species			X		X		X		0.2	1.0	5.1	0.0	-	28.8	-	-	40.0	-	-	30.1	-	-	29.0

*Overall observed seroprevalence in 2013 field study was 11.1% in urban, 15.3% in peri-urban, and 24.0% in rural areas. Predicted seroprevalence were only calculated for animal exposure scenarios reported by >3% of at least one subgroup; "-" indicates scenarios where predicted seroprevalence were not calculated.

Table 8. The most common community-level exposure to animal species in each residential setting. Exposure was defined as the presence of the animal species at the individual's community. BNs RC-A, RC-B and RC-C were used to predict seroprevalence under each of the scenarios shown below, and summarised in Figure 5b.

Section	Animal species present in community				% of population exposed to animal species			Observed seroprevalence* (%)			Predicted seroprevalence using RC-A (%)			Predicted seroprevalence using RC-B (%)			Predicted seroprevalence using RC-C (%)		
	Cow	Goat	Horse	Pig	Urban n=579	Peri-urban n=287	Rural n=1286	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural
A. Exposure scenarios related to EACH animal species	X				9.3	14.3	30.0	24.1	22.0	25.1	14.5	19.9	30.2	24.0	21.9	25.2	24.0	21.9	25.2
		X			7.1	8.0	13.8	29.3	17.4	25.3	14.8	20.3	30.7	29.2	17.3	25.3	29.2	17.3	25.3
			X		8.1	7.0	24.1	25.5	30.0	25.2	15.0	20.5	31.0	25.5	29.9	25.2	25.5	29.9	25.2
				X	14.5	18.8	32.9	23.8	25.9	27.4	15.9	21.6	32.5	23.8	25.8	27.5	23.8	25.8	27.5
B. Exposure scenarios related to combinations of TWO animal species	X	X			6.7	7.3	11.7	30.8	19.1	28.0	19.3	25.8	37.8	51.3	24.5	26.5	35.5	13.8	28.8
	X		X		7.8	6.3	18.4	26.7	22.2	24.2	19.5	26.1	38.1	46.6	39.8	26.4	32.2	23.7	26.3
	X			X	7.6	7.3	20.1	27.3	23.8	27.5	20.5	27.4	39.7	44.3	35.0	28.7	27.2	23.7	27.6
		X	X		6.6	5.2	9.8	31.6	26.7	27.0	19.9	26.5	38.7	53.2	33.1	26.5	37.0	17.9	28.6
		X		X	6.6	3.8	8.9	31.6	9.1	30.7	20.9	27.9	40.3	50.9	28.8	28.9	31.5	9.06	30.7
			X	X	7.3	3.8	16.7	28.6	18.2	27.0	21.2	28.1	40.6	46.2	45.1	28.7	28.5	18.1	27.0
C. Exposure scenarios related to combinations of THREE animal species	X	X	X		6.6	5.2	9.4	31.6	26.7	26.5	25.3	33.1	46.4	74.3	43.3	27.8	41.4	10.2	29.9
	X	X		X	6.6	3.5	8.4	31.6	10.0	31.5	26.6	34.6	48.0	72.5	38.4	30.2	35.6	8.2	30.9
	X		X	X	7.1	3.5	13.6	29.3	10.0	26.3	26.9	34.9	48.3	68.6	56.0	30.0	32.4	16.5	27.1
		X	X	X	6.4	3.1	7.5	32.4	11.1	30.2	27.4	35.4	49.0	74.1	48.8	30.2	37.1	6.0	30.3

*Overall observed seroprevalence in 2013 field study was 11.1% in urban, 15.3% in peri-urban, and 24.0% in rural areas. Predicted seroprevalence were only calculated for animal exposure scenarios reported by >3% of at least one subgroup; "-" indicates scenarios where predicted seroprevalence were not calculated.

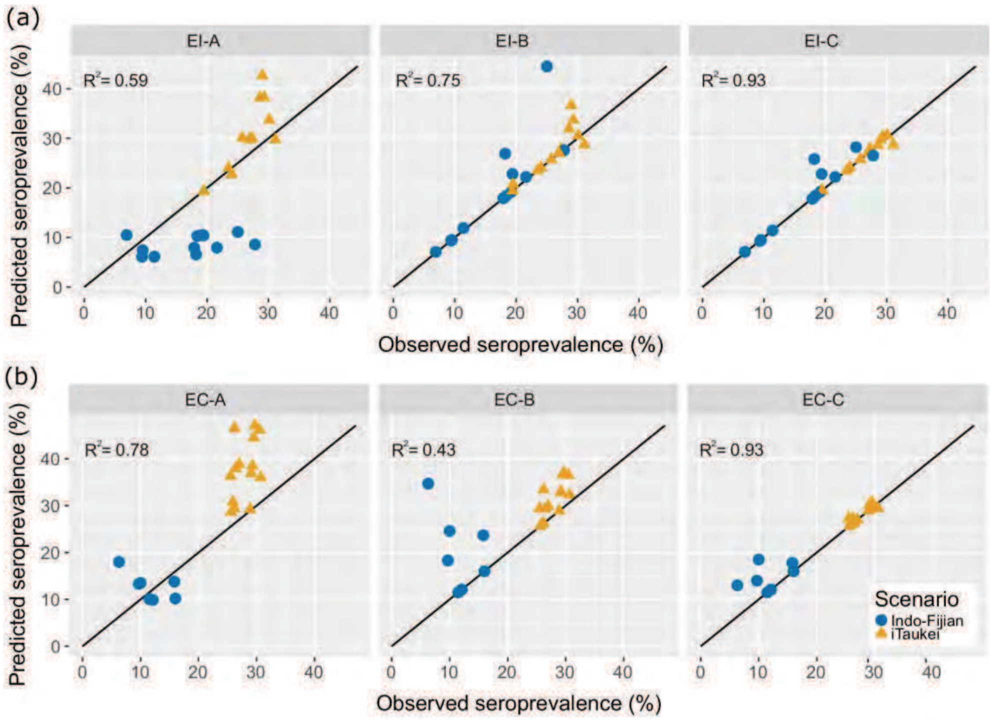

Figure 4. a) Comparison between observed and predicted seroprevalence using Bayesian networks EI-A, EI-B, and EI-C models for individual/household-level exposure for each ethnic group. b) Comparison between observed and predicted seroprevalence using Bayesian networks EC-A, EC-B, and EC-C models for community-level exposure for each ethnic group.

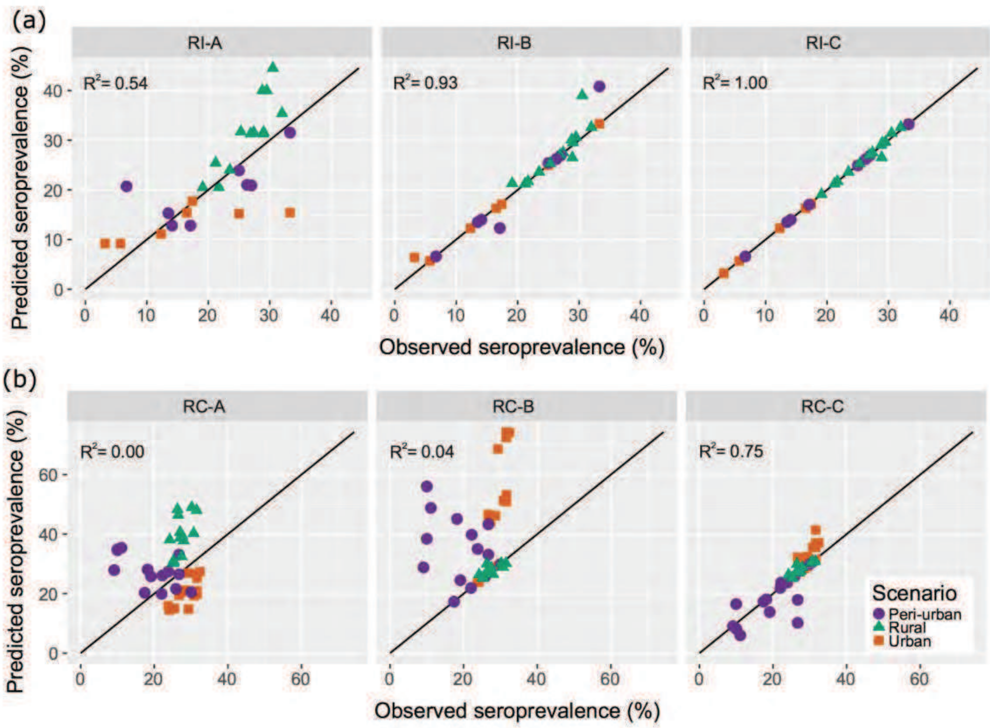

Figure 5. a) Comparison between observed and predicted seroprevalence using Bayesian networks RI-A, RI-B, and RI-C models for individual/household-level exposure and each residential setting. b) Comparison between observed and predicted seroprevalence using Bayesian networks RC-A, RC-B, and RC-C models for community-level exposure and each residential setting.

Relative importance of animal species under different exposure scenarios

Group C BNs showed the best predictive performance, and were used to determine the relative importance of animal species under different scenarios of ethnicity and residential setting. Table 9 shows results of scenario analyses for individual/household-level exposures in ethnic groups (BN EI-C). The prevalence of animal exposures differed markedly between the two ethnic groups, and the animal species associated with higher seroprevalence also varied. For example, 12.2% of Indo-Fijians owned goats, and this scenario was associated with a higher seroprevalence of 17.8% compared to Indo-Fijians who do not own goats (6.0%, p=0.002). Only 3.1% of iTaukei owned goats, but this ethnic group was more likely to report physical contact with rodents (17.3%), and this exposure was associated with higher seroprevalence (27.9%) compared to those who do not have contact with rodents (21.6%, p=0.021). Figure 6a highlights differences in individual/household animal exposure between ethnic groups, and relative importance of each species on seroprevalence. Triangles and circles represent statistically significant or insignificant differences in seroprevalence between exposed and un-exposed groups.

Table 10 shows the results of scenario analyses for community-level exposures in ethnic groups (BN EC-C). The most common livestock animals found in iTaukei communities were pigs (32.7%) and cows (24.8%). Many communities had multiple livestock species, e.g.13.4% of iTaukei communities reported the presence of cows *and* pigs *and* horses, and this scenario was associated with a higher predicted seroprevalence of 27.3% compared to communities without any of those animal species (20.6%, p=0.030). In contrast, the most common livestock in Indo-Fijian communities were cows (12.6%) and goats (11.3%). Although only 8.7% of Indo-Fijian communities reported the presence of two or more livestock species, the presence of cows *and* horses (reported by 4.4% of Indo-Fijians) was associated with a higher predicted seroprevalence of 18.5% compared to 6.3% in those who were not exposed to these species (p=0.036). Figure 6b highlights the differences in exposure and relative importance of animal exposures between ethnic groups.

Table 9. Difference in seroprevalence based on ethnicity and individual/household-level exposure to animal species or combinations of species. BN EI-C was used to predict seroprevalence in exposed and unexposed groups. Results for individual species are summarized in Figure 6a.

Physical contact		Animal species present at household						% of population exposed to animal species		Predicted seroprevalence in exposed (%)		Predicted seroprevalence in unexposed (%)		*p* value for statistical difference in seroprevalence between exposed and unexposed#	
Rodents	**Mongoose**	**Dog**	**Cat**	**Cow**	**Goat**	**Horse**	**Pig**	**iTaukei n=1651**	**Indo-Fijian n=459**	**iTaukei**	**Indo-Fijian**	**iTaukei**	**Indo-Fijian**	**iTaukei**	**Indo-Fijian**
X								17.3	6.3	27.9	7.1	21.6	7.4	**0.021**	0.953
	X							7.5	2.0	30.6	-	22.0	-	**0.029**	-
		X						26.0	43.6	24.1	9.5	22.2	5.8	0.419	0.135
			X					14.4	23.3	19.4	9.3	23.2	6.9	0.196	0.408
				X				14.1	10.7	27.1	18.4	22.0	6.1	0.085	**0.002**
					X			3.1	12.2	23.5	17.8	22.7	6.0	0.893	**0.002**
						X		10.5	5.7	27.0	19.2	22.2	6.7	0.153	**0.018**
							X	13.7	0.4	25.7	-	22.2	-	0.243	-
X	X							4.6	1.1	29.7	-	21.2	-	0.081	-
		X	X					6.8	17.2	19.5	11.4	22.5	6.1	0.465	0.122
		X			X			1.5	8.1	-	22.2	-	4.6	-	**<0.001**
		X		X				6.2	7.8	28.7	22.8	21.5	4.8	0.092	**<0.001**
				X		X		7.2	2.6	28.6	-	21.9	-	0.093	-
		X	X		X			0.7	4.8	-	25.8	-	4.9	**-**	**<0.001**
		X	X	X				1.8	3.9	-	26.5	-	5.0	**-**	**<0.001**
		X		X	X			1.3	4.4	-	28.2	-	4.3	**-**	**<0.001**
		X		X		X		3.5	2.6	30.3	-	21.4	-	0.110	-

*Overall observed seroprevalence in 2013 field study was 22.7% in iTaukei and 7.4% in Indo-Fijians. #Using test of difference between proportions, statistically significant results ($p<0.05$) in bold.

Table 10. Difference in seroprevalence based on ethnicity and community-level exposure to animal species or combinations of species. BN EC-C was used to predict seroprevalence in exposed and unexposed groups. Results for individual species are summarized in Figure 6b.

Animal species present in community				% of population exposed to animal species		Predicted seroprevalence in exposed (%)		Predicted seroprevalence in unexposed (%)		*p* value for statistical difference in seroprevalence between exposed and unexposed#	
Cow	**Goat**	**Horse**	**Pig**	**iTaukei n=1651**	**Indo-Fijian n=459**	**iTaukei**	**Indo-Fijian**	**iTaukei**	**Indo-Fijian**	**iTaukei**	**Indo-Fijian**
X				24.8	12.6	25.6	12.1	21.7	6.7	0.102	0.142
	X			11.3	11.3	28.9	11.5	21.9	6.9	**0.031**	0.233
		X		21.1	5.4	26.2	16.0	21.7	6.9	0.075	0.091
			X	32.7	1.3	26.1	-	21.0	-	**0.020**	-
X	X			10.7	6.8	29.7	14.0	21.2	6.2	**0.011**	0.097
X		X		16.7	4.4	27.3	18.5	21.1	6.3	**0.026**	**0.036**
X			X	18.6	0.9	26.4	-	20.8	-	**0.038**	-
	X	X		9.6	4.1	30.3	17.8	21.2	6.4	**0.009**	0.056
	X		X	9.5	0.7	29.3	-	20.9	-	**0.018**	-
		X	X	15.9	0.4	27.0	-	20.8	-	**0.031**	-
X	X	X		9.4	3.5	30.7	13.0	20.7	5.7	**0.005**	0.228
X	X		X	9.1	0.4	29.7	-	20.6	-	**0.012**	-
X		X	X	13.4	0.4	27.3	-	20.6	-	**0.030**	-
	X	X	X	8.4	0.4	30.3	-	20.7	-	**0.010**	-

*Overall observed seroprevalence in 2013 field study was 22.7% in iTaukei and 7.4% in Indo-Fijians.

#Using test of difference between proportions, statistically significant results ($p<0.05$) in bold.

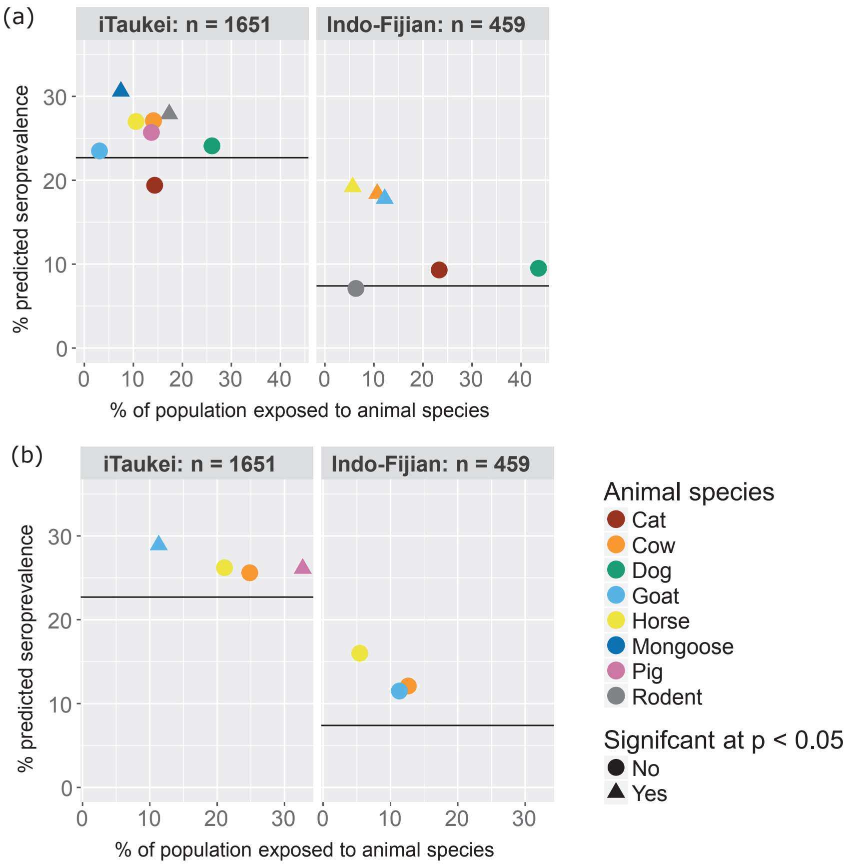

Figure 6. a) Individual/household-level exposure to animals – differences in exposure and predicted seroprevalence between ethnic groups. Exposure is defined as physical contact with rodents or mongoose, or presence of other animal species at the individual's household. b) Community-level exposure to animals – differences in exposure and predicted seroprevalence between ethnic groups. Exposure is defined as the presence of animal species at the individual's community. Horizontal black lines indicate mean seroprevalence for each subgroup. Triangles/circles indicate statistically significant/insignificant difference in seroprevalence between exposed and un-exposed groups.

Table 11 shows the results of scenario analyses for individual/household-level exposures in different residential settings (BN RI-C). In urban areas, the most common animal exposures were to dogs (26.6%), cats (15.2%), and rodents (13.6%). Few urban residents reported exposure to cows (3.5%) or pigs (3.6%), but their presence at households was associated with a higher predicted seroprevalences of 25.0% (vs 10.6%, p=0.044) and 33.3% (vs 10.2%, p<0.001) compared to those without these exposures. In rural areas, physical contact with rodents (16.1%) and mongoose (7.8%) were more common than in urban or peri-urban areas, and associated with higher seroprevalence of 29.5% (vs 22.9%, p=0.042) and 32.6% (vs 23.3%, p=0.037). Figure 7a highlights the differences in exposure and relative importance of individual/household-level animal exposures between urban, peri-urban, and rural areas.

Table 12 provides results of scenario analyses for community-level exposures in residential settings (BN RC-C). Pigs were the most common livestock species in all community types, present in 14.5% of urban, 18.8% of peri-urban, and 32.9% of rural communities. Pigs were associated with higher seroprevalence in all community types, but particularly striking in urban areas where exposure was associated with a seroprevalence of 23.8%, compared to 8.9% in urban dwellers who were not exposed to pigs (p<0.001). Multiple livestock species in urban areas was associated with very high predicted seroprevalence, e.g. 35.6% in urban communities with cows *and* goats *and* pigs (p<0.001). Figure 7b highlights the relative importance of animal species in each residential setting.

Table 11. Difference in seroprevalence based on residential setting and individual/household-level exposure to animal species or combinations of species. BN RI-C was used to predict seroprevalence in exposed and unexposed groups. Results for individual species are summarized in Figure 7a.

Physical contact		Animal species present at household						% of population exposed to animal species			Predicted seroprevalence in exposed (%)			Predicted seroprevalence in unexposed (%)			*p* value for statistical difference in seroprevalence between exposed and unexposed#		
Rodents	Mongoose	Dog	Cat	Cow	Goat	Horse	Pig	Urban n=579	Peri-urban n=287	Rural n=1286	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural
X								13.6	12.9	16.1	16.3	26.9	29.5	10.2	13.6	22.9	0.108	**0.036**	**0.042**
	X							4.0	4.2	7.8	17.2	24.9	32.6	10.8	14.9	23.3	0.338	0.347	**0.037**
		X						26.6	33.4	30.7	12.3	13.5	23.6	10.6	16.3	24.2	0.564	0.535	0.816
			X					15.2	22.3	15.8	5.7	14.0	21.7	12.0	15.7	24.5	0.083	0.739	0.392
				X				3.5	5.2	19.4	25.0	6.6	27.0	10.6	15.8	23.3	**0.044**	0.336	0.220
					X			0.9	1.0	7.7	-	-	21.2	-	-	24.3	-	-	0.488
						X		1.4	2.1	14.5	-	-	27.5	-	-	23.4	-	-	0.226
							X	3.6	6.6	14.8	33.3	26.2	25.3	10.2	14.6	23.8	**<0.001**	0.176	0.655
X	X							2.8	3.1	4.6	-	33.2	31.4	-	0	22.5	-	**<0.001**	0.115
		X	X					5.4	14.3	9.8	3.2	17.0	19.1	11.2	17.3	24.1	0.164	0.964	0.218
		X		X				0.7	2.4	10.0	-	-	26.5	-	-	23.5	-	-	0.461
				X		X		0.7	1.7	9.5	-	-	29.6	-	-	23.3	-	-	0.125
		X		X		X		0.2	1.0	5.1	-	-	29.0	-	-	23.5	-	-	0.316

*Overall observed seroprevalence in 2013 field study was 11.1% in urban, 15.3% in peri-urban, and 24.0% in rural areas.

#Using test of difference between proportions, statistically significant results ($p<0.05$) in bold.

Table 12. Difference in seroprevalence based on residential setting and community-level exposure to animal species or combinations of species. BN RC-C was used to predict seroprevalence in exposed and unexposed groups. Results for individual species are summarized in Figure 7b.

Animal species present in community				% of population exposed to animal species			Predicted seroprevalence in exposed (%)			Predicted seroprevalence in unexposed (%)			*p* value for statistical difference in seroprevalence between exposed and unexposed#		
Cow	Goat	Horse	Pig	Urban n=579	Peri-urban n=287	Rural n=1286	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural
X				9.3	14.3	30.0	24.0	21.9	25.2	9.7	14.2	23.5	**0.001**	0.205	0.513
	X			7.1	8.0	13.8	29.2	17.3	25.3	9.7	15.2	23.8	**0.001**	0.789	0.664
		X		8.1	7.0	24.1	25.5	29.9	25.2	9.8	14.2	23.7	**0.001**	0.060	0.590
			X	14.5	18.8	32.9	23.8	25.8	27.5	8.9	12.9	22.3	**<0.001**	**0.018**	**0.040**
X	X			6.7	7.3	11.7	35.5	13.8	28.8	9.2	13.9	23.7	**<0.001**	0.990	0.180
X		X		7.8	6.3	18.4	32.2	23.7	26.3	9.2	13.1	23.4	**<0.001**	0.209	0.358
X			X	7.6	7.3	20.1	27.2	23.7	27.6	8.9	12.2	22.7	**<0.001**	0.138	0.113
	X	X		6.6	5.2	9.8	37.0	17.9	28.6	9.3	14.0	23.7	**<0.001**	0.674	0.229
	X		X	6.6	3.8	8.9	31.5	9.06	30.7	9.0	12.2	22.9	**<0.001**	0.755	0.068
		X	X	7.3	3.8	16.7	28.5	18.1	27.0	9.0	11.6	22.5	**<0.001**	0.516	0.169
X	X	X		6.6	5.2	9.4	41.4	10.2	29.9	9.1	12.7	23.7	**<0.001**	0.777	0.139
X	X		X	6.6	3.5	8.4	35.6	8.2	30.9	8.9	11.6	23.2	**<0.001**	0.741	0.082
X		X	X	7.1	3.5	13.6	32.4	16.5	27.1	9.0	11.0	22.8	**<0.001**	0.591	0.231
	X	X	X	6.4	3.1	7.5	37.1	6.0	30.3	9.0	11.0	23.0	**<0.001**	0.636	0.115

*Overall observed seroprevalence in 2013 field study was 11.1% in urban, 15.3% in peri-urban, and 24.0% in rural areas.

#Using test of difference between proportions, statistically significant results (p<0.05) in bold.

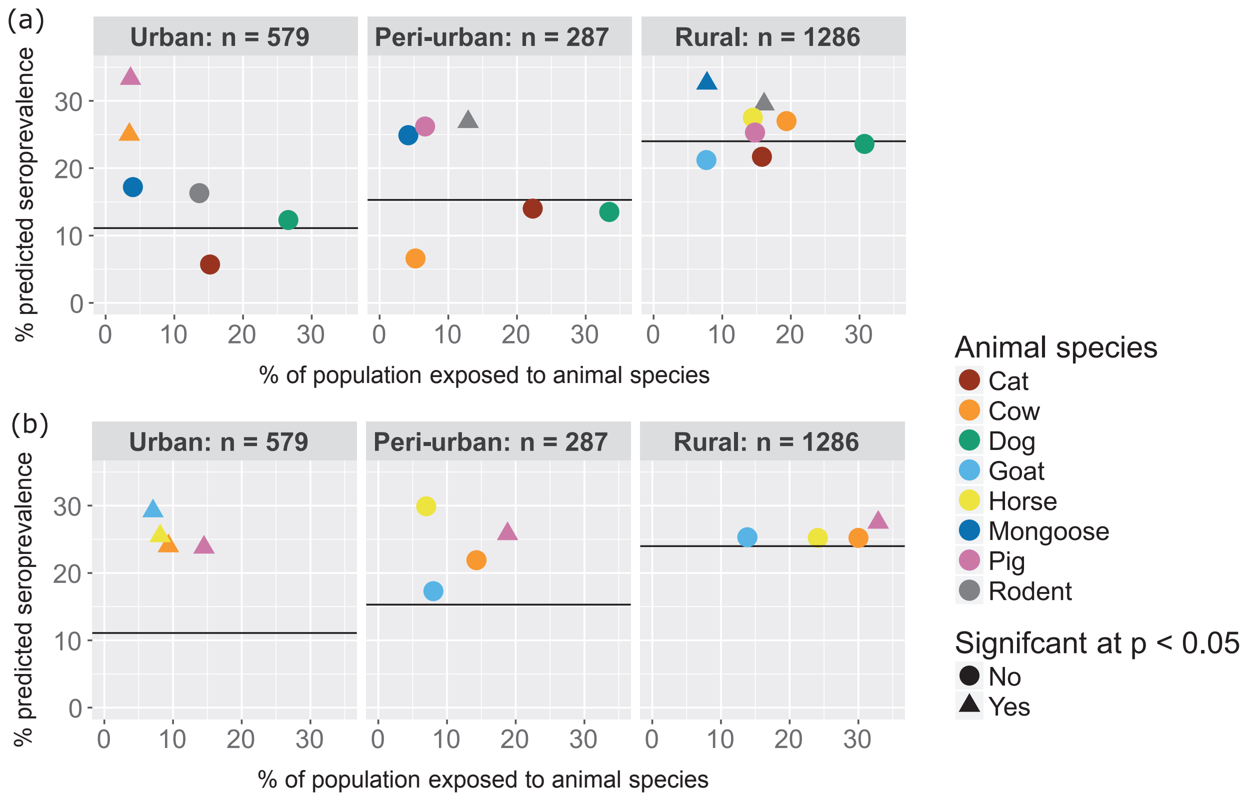

Figure 7. a) Individual/household-level exposure to animals – differences in exposure and predicted seroprevalence between residential settings. Exposure is defined as the presence of animal species at the individual's household. b) Community-level exposure to animals – differences in exposure and predicted seroprevalence between residential settings. Exposure is defined as the presence of animal species at the individual's community. Horizontal black lines indicate mean seroprevalence for each subgroup. Triangles/circles indicate a statistically significant/insignificant difference in seroprevalence between exposed and un-exposed groups.

DISCUSSION

Our study showed that model performance of BNs can be significantly improved by building models that better represent causality and account for dependencies among predictor and indicator variables. Group C BNs were structured to better represent causality and complex interdependencies between nodes, and performed better than the naïve BNs in Group A that were conceptually similar to standard regression models (i.e. predictor/indicator variables were independent). Our study demonstrated some useful features of BNs, including the ability to refine model structure, include strongly correlated predictor and indicator variables, and predict outcomes under complex scenarios. We used leptospirosis as a case study, but the approaches presented here could potentially be used to model other diseases or health outcomes.

We determined AUC for our models and obtained values between 0.58 and 0.63 for 25th and 75th percentile limits (Table 8), indicating poor model discrimination power. This means the probability distributions for true positives and true negatives overlapped significantly, which can occur in situations of low prevalence, where the model never predicts a high probability for true cases. This could explain the poor AUC results for our models.[37] Also, AUC does not take into account the predicted probability values and model goodness-of-fit.[37] There are many other metrics that are commonly used to assess the performance of presence/absence models such as the ones presented in this paper. These include sensitivity (true positive rate), specificity (true negative rate), the True Skill Statistic (TSS) and Kappa Statistic. We chose to use AUC in our study because it measures performance across multiple cutoffs, while TSS and Kappa Statistic use the most probable outcome as the cutoff. The limitation of using the most probable outcome as the cutoff is that the metric becomes less reliable when prevalence rates (true positive rates) for the outcome being predicted are low, because in these situations a model trained on data containing low prevalence rates will rarely predict a high probability of presence for any scenario. In our study, the predicted and actual seroprevalence of *Leptospira* for any scenario were mostly below 30%, indicating low prevalence rates.

We compared predicted versus observed seroprevalence to assess the predictive performance of our BNs and found that Group C models (with the most complex structures) had the best performance, with R^2 values of 0.75 to 1.00, and lower MSE compared to Group A and B models. Using the Group C models, we found that scenario analyses provided important insights into the relative importance of animal species in leptospirosis transmission in different ethnic groups and residential settings. These insights were gained by predicting outcomes under complex scenarios that included multiple correlated predictor/indicator variables, which would have been more difficult to achieve with regression methods. A central challenge in leptospirosis control is to identify specific points in exposure pathways where public health and environmental health interventions are likely to be most effective. Because of the complex and variable transmission dynamics of leptospirosis, scenario analyses using BNs could be useful for providing insights to inform more targeted prevention and control strategies for subpopulations.

In the Pacific Islands, leptospirosis has been identified in many animal species including rodents and livestock,[38] and is considered as one of the most important livestock diseases in the region in terms of impact on human health.[39] However, the relative importance of each animal disease in human infections is currently poorly understood. Our results provide epidemiological evidence that multiple animal species are likely to be important in leptospirosis transmission in Fiji, and that the intensity of exposure to animals as well as the relative importance of each animal species vary significantly between ethnic groups and residential settings, i.e. the prevalence of risk factors vary significantly between subgroups. For iTaukei, contact with rodents and mongoose and community-level exposure to livestock were strongly associated with infection. In contrast, very few Indo-Fijians reported contact with rodents or mongoose, but household exposure to livestock was important. In rural areas, physical contact with rodents and mongoose were important. Community-level exposure to pigs was important in all residential settings. Importantly, in urban settings, exposure to livestock was associated with a very high risk of infection, possibly because animals are kept closer to homes compared to rural areas. This finding is concerning, because the combination of population growth, urbanisation, and agricultural intensification (including subsistence farming) might fuel future urban outbreaks in this setting.

Detailed insights about the role of different animal species in different socio-ecological niches could potentially be useful for designing interventions that are specifically relevant for subgroups, e.g. health promotion messages related to contact with rodents and mongoose should be particularly strengthened in iTaukei communities, but improving management of livestock animals is important for all communities in Fiji. Animal and anthropological studies will be required to confirm the epidemiological associations identified by our study. Our findings provide important baseline data for developing future studies to assess the impact of interventions in Fiji, e.g. evaluating specific strategies for each ethnic group and residential setting.

Our results should be interpreted in light of the study's limitations. The study's outcome measure was the presence of *Leptospira* antibodies, which is an indication of prior infection. However, many leptospirosis infections are asymptomatic and the severity of clinical infections depends on many factors including age, comorbidities, and pathogen virulence. Our study used animal data at the place of residence, but it is possible for infections to occur elsewhere. The database was obtained from a cross-sectional study conducted in 2013, and it is possible for risk factors to evolve over time.

The application of BNs in infectious disease epidemiology has recently been increasing. A recent study used BNs to model meningitis outbreaks in the Niger using historical epidemiological databases, and concluded that BNs provide a promising approach for understanding the dynamics of epidemics, estimating the risk of outbreaks, and providing information to target control interventions.[40] BNs have also been used to model seasonal and population influences on pneumonic plague,[41] the impact of demographics and vaccination on influenza,[42] hierarchical relationships of risk factors associated with infectious diarrhoea in children,[43] and household factors that influence the risk of malaria in sub-Saharan Africa.[44]

There are other advantages of Bayesian networks that were not fully explored in this study, including their graphical user interface that allow models to be more easily understood and interpreted; the interactive and dynamic setup that allows users to define complex scenarios and see updated predictions almost immediately; the ability to incorporate different sources and types of knowledge including empirical data and expert opinion; the ease with which new data can be incorporated into models to update probabilistic relationships between variables; the ability to model causal pathways; and the ability to use the models in predictive or diagnostic modes, or a combination of both as shown in the BNs used in this study.[11,12,33] BNs have therefore been used in many disciplines including medicine, ecology, environmental sciences, engineering, gaming, and artificial intelligence.

Future work on the use of BNs in infectious disease epidemiology should explore the development of more complex models that incorporate a wider range or predictor/indicator variables, including variables that operate at different ecological scales. Integrating BNs with other types of models that include spatial, temporal or dynamic components will also help improve understanding of disease transmission.

CONCLUSIONS

We demonstrated that BNs provide a promising alternative approach to modelling infectious disease epidemiology and unravelling the complex drivers of transmission. Using BNs, our study provided important information on the role of different animal species in leptospirosis transmission in Fiji. We showed that the drivers of leptospirosis transmission are likely to vary between socio-ecological niches, with important implications for targeted prevention and control strategies. Although our study focused on leptospirosis in Fiji, the analytical approaches could be used to model other diseases or health outcomes.

AUTHOR CONTRIBUTIONS

CL and CS conceived the study and proposed the use of Bayesian networks; CL, HM, and CS designed the study, analysed the data, and developed the figures and tables. CW, CL, MK, and EN contributed to the design of the field study and collection of survey data. JL, CL, and CW integrated spatial environmental and socio-demographic data with the survey data. CL, HM, and CS prepared the draft manuscript. All authors reviewed and approved the final manuscript.

ACKNOWLEDGEMENTS

We extend our warmest thanks to the many participants, communities, Ministry of Health staff and community health workers who generously contributed to the field study conducted in Fiji in 2013. We also thank the WHO/FAO/OIE Collaborating Centre for Reference and Research on Leptospirosis (at Health Support Queensland, Department of Health, Brisbane, Australia) for providing the laboratory support for the study.

CONFLICTS OF INTEREST:

The authors do not have any conflicts of interest to declare.

FUNDING SOURCES:

CLL was supported by an Australian National Health and Medical Research Council (NHMRC) Fellowship (1109035). CHW was supported by the UK Medical Research Council (grant MR/J003999/1) and the Chadwick Trust. The study was partly funded by a University of Queensland Early Career Researcher Grant (2014003059). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Eisenberg JN, et al. Environmental Determinants of Infectious Disease: A Framework for Tracking Causal Links and Guiding Public Health Research. *Environ Health Perspect.* 2007;115:1216-23.

2. Rabinowitz PM, et al. Toward Proof of Concept of a One Health Approach to Disease Prediction and Control. *Emerg Infect Dis.* 2013;19:e130265.

3. Mills JN, et al. Potential Influence of Climate Change on Vector-Borne and Zoonotic Diseases: A Review and Proposed Research Plan. *Environ Health Perspect.* 2010;118:1507-14.

4. Jones KE, et al. Global Trends in Emerging Infectious Diseases. *Nature.* 2008;451:990-3.

5. McMichael AJ. Prisoners of the Proximate: Loosening the Constraints on Epidemiology in an Age of Change. *Am J Epidemiol.* 1999;149:887-97.

6. Susser M. Does Risk Factor Epidemiology Put Epidemiology at Risk? Peering into the Future. *J Epidemiol Community Health.* 1998;52:608-11.

7. Burns CJ, et al. Evaluating Uncertainty to Strengthen Epidemiologic Data for Use in Human Health Risk Assessments. *Environ Health Perspect.* 2014;122:1160-5.

8. Lau CL, et al. Bayesian Networks in Infectious Disease Eco-Epidemiology. *Rev Environ Health.* 2016.

9. Keyes KM, et al. Commentary: The Limits of Risk Factors Revisited: Is It Time for a Causal Architecture Approach? *Epidemiology.* 2017;28:1-5.

10. Poole C. Commentary: Some Thoughts on Consequential Epidemiology and Causal Architecture. *Epidemiology.* 2017;28:6-11.

11. Joffe M, et al. Causal Diagrams in Systems Epidemiology. *Emerg Themes Epidemiol.* 2012;9:1.

12. Landuyt D, et al. A Review of Bayesian Belief Networks in Ecosystem Service Modelling. *Environ Model Softw.* 2013;46:1-11.

13. Costa F, et al. Global Morbidity and Mortality of Leptospirosis: A Systematic Review. *PLoS Negl Trop Dis.* 2015;9:e0003898.

14. Bharti AR, et al. Leptospirosis: A Zoonotic Disease of Global Importance. *Lancet Infect Dis.* 2003;3:757-71.

15. Lau CL, et al. Climate Change, Flooding, Urbanisation and Leptospirosis: Fuelling the Fire? *Trans R Soc Trop Med Hyg.* 2010;104:631-8.

16. Mwachui MA, et al. Environmental and Behavioural Determinants of Leptospirosis Transmission: A Systematic Review. *PLoS Negl Trop Dis.* 2015;9:e0003843.

17. Jones BA, et al. Zoonosis Emergence Linked to Agricultural Intensification and Environmental Change. *Proc Natl Acad Sci.* 2013;110:8399-404.

18. Hartskeerl RA, et al. Emergence, Control and Re-Emerging Leptospirosis: Dynamics of Infection in the Changing World. *Clin Microbiol Infect.* 2011;17:494-501.

19. Coker R, et al. Towards a Conceptual Framework to Support One-Health Research for Policy on Emerging Zoonoses. *Lancet Infect Dis.* 2011;11:326-31.

20. Liverani M, et al. Understanding and Managing Zoonotic Risk in the New Livestock Industries. *Environ Health Perspect.* 2013;121:873-7.

21. McIver L, et al. Health Impacts of Climate Change in Pacific Island Countries: A Regional Assessment of Vulnerabilities and Adaptation Priorities. *Environ Health Perspect.* 2015.

22. Averett N. Pacific Island Countries and Climate Change: Examining Associated Human Health Vulnerabilities. *Environ Health Perspect.* 2016;124:A208.

23. Kline K, et al. Neglected Tropical Diseases of Oceania: Review of Their Prevalence, Distribution, and Opportunities for Control. *PLoS Negl Trop Dis.* 2013;7:e1755.

24. Weinberger D, et al. El Nino Southern Oscillation and Leptospirosis Outbreaks in New Caledonia. *PLoS Negl Trop Dis.* 2014;8:e2798.

25. Massenet D, et al. An Unprecedented High Incidence of Leptospirosis in Futuna, South Pacific, 2004 - 2014, Evidenced by Retrospective Analysis of Surveillance Data. *PLoS One.* 2015;10:e0142063.

26. Colt S, et al. Human Leptospirosis in the Federated States of Micronesia: A Hospital-Based Febrile Illness Survey. *BMC Infect Dis.* 2014;14:186.

27. Lau C, et al. Human Leptospirosis Infection in Fiji: An Eco-Epidemiological Approach to Identifying Risk Factors and Environmental Drivers for Transmission. *PLoS Negl Trop Dis.* 2016;10:e0004405.

28. Lau CL, et al. Leptospirosis in American Samoa - Estimating and Mapping Risk Using Environmental Data. *PLoS Neglected Tropical Diseases.* 2012;6:e1669.

29. Ram P, et al. Further Observations on the Epidemiology of Leptospirosis in Fiji. *Fiji Medical Journal.* 1982;10:71-5.

30. Ghosh A, et al. Leptospirosis in Fiji: Incidence between 2000 to 2007 and Review of Literature. *Fiji Medical Journal.* 2010;29:8-14.

31. McIver L, et al. Climate Change and Health in Fiji: Environmental Epidemiology of Infectious Diseases & Potential for Climate-Based Early Warning Systems. *Fiji Journal of Public Health.* 2012;1:7-13.

32. Fiji Bureau of Statistics. Census of Population and Housing. 2007. http://www.statsfiji.gov.fj/index.php/2007-census-of-population. Accessed 8 September 2015.

33. Fenton N, et al. Risk Assessment and Decision Analysis with Bayesian Networks. Boca Raton: CRC Press, Taylor & Francis Group.; 2013.

34. Norsys Software Corp. Netica Bayesian Belief Network Software. https://http://www.norsys.com2013.

35. Fenton N, et al. Chapter 7. Defining the Structure of Bayesian Networks. Risk Assessment and Decision Analysis with Bayesian Networks. Boca Raton: CRC Press, Taylor & Francis Group.; 2013.

36. Do CB, et al. What Is the Expectation Maximization Algorithm? *Nat Biotechnol.* 2008;26:897-9.

37. Lobo JM, et al. Auc: A Misleading Measure of the Performance of Predictive Distribution Models. *Global Ecology and Biogeography.* 2007;17:145-51.

38. Brioudes A, et al. A Review of Domestic Animal Diseases within the Pacific Islands Region. *Acta Trop.* 2014;132:23-38.

39. Brioudes A, et al. Diseases of Livestock in the Pacific Islands Region: Setting Priorities for Food Animal Biosecurity. *Acta Trop.* 2015;143:66-76.

40. Beresniak A, et al. A Bayesian Network Approach to the Study of Historical Epidemiological Databases: Modelling Meningitis Outbreaks in the Niger. *Bull World Health Organ.* 2012;90:412-7A.

41. Dawson P, et al. Detecting Disease Outbreaks Using a Combined Bayesian Network and Particle Filter Approach. *Journal of Theoretical Biology.* 2015;370:171-83.

42. Abbas K, et al. Temporal Analysis of Infectious Diseases: Influenza. Proceedings of the 2005 ACM symposium on Applied computing. Santa Fe, New Mexico: ACM; 2005:267-71.

43. Nguefack-Tsague G. Using Bayesian Networks to Model Hierarchical Relationships in Epidemiological Studies. *Epidemiology and Health.* 2011;33:e2011006.

44. Semakula HM, et al. A Bayesian Belief Network Modelling of Household Factors Influencing the Risk of Malaria: A Study of Parasitaemia in Children under Five Years of Age in Sub-Saharan Africa. *Environmental Modelling & Software.* 2016;75:59-67.